SEGMENTATION-BASED MONGOLIAN LVCSR APPROACH

Feilong Bao, Guanglai Gao, Xueliang Yan, Weihua Wang

College of Computer Science, Inner Mongolia University,
Hohhot 010021, China
{csfeilong, csggl, csyxl}@imu.edu.cn wangweihua23@gmail.com

ABSTRACT

Mongolian is an agglutinative language. Each root can be followed by several suffixes to formulate new words. This special word formation characteristic results in probably millions of Mongolian words, which is far beyond the coverage of the pronunciation dictionary of any current Mongolian speech recognition system. Moreover, even if the pronunciation dictionary is large enough to cover all of the Mongolian words, the recognition system still cannot perform well due to the problem of sample sparseness. In this paper, we propose a segmentation-based Mongolian Large Vocabulary Continuous Speech Recognition (LVCSR) approach and rebuild the corresponding acoustic model and language model. Experimental results show that, by converting most of these words into their corresponding In-Vocabulary form, the proposed approach effectively recognizes most of the Mongolian words and greatly improves the sample sparseness problem in the language model.

Index Terms— Mongolian, segmentation, stem, ending suffix, LVCSR

1. INTRODUCTION

The research of the Traditional Mongolian speech recognition technology starts at the beginning of the 21st century and a lot of work has been done by Gao et al. [1] and Hasi et al. [2]. Based on their work, Bao et al. [3] improved the acoustic model and built a system with good performance on a testing set that contains about 10000 common words.

Mongolian is an agglutinative language [4]. Most words are composed of a root followed by several suffixes. This word formation characteristic results in probably millions of words, which is far beyond the coverage of the pronunciation dictionary of any current Mongolian LVCSR system. Enlarging the scale of the pronunciation dictionary is, of course, a possible solution. But large pronunciation dictionary means much more training data are required to train the language model. However, the rare corpus in text form available makes this scheme not practical. It's high time to seek new approaches which do not need to endlessly enlarge the scale of the pronunciation dictionary.

Based on the characteristic of Mongolian word formation, we proposed a segmentation-based LVCSR approach which effectively solved the OOV words recognition problem and reduce the requirement for training data when estimating the language model.

The rest of the paper is organized as follows: section 2 describes the characteristic of Mongolian word formation in detail; section 3 depicts the segmentation-based Mongolian LVCSR approach; experimental setup and results are shown in section 4; finally, the conclusions are summarized in section 5.

2. CHARACTERISTIC OF MONGOLIAN WORD FORMATION

In general, a Mongolian word can be decomposed into two parts: a root and several suffixes. There is no prefix in Mongolian language and the suffixes can be categorized as word-formation suffix, inflectional suffix and ending suffix. Generally, the ending suffixes affect only the form, such as tense, person, etc. Every root can be followed by multiple word-formation suffixes and inflectional suffixes to generate the stem and the generated stem can further be followed by one ending suffix. Moreover, the pronunciation of the suffixes will be different depending on the stems they are following. Figure 1 illustrates the relationships among the root, stem and suffixes of a Mongolian word. According to our statistics, the number of stems that are commonly used in Mongolian is about 30000 and that of ending suffixes is 400 more or less. Nearly one million words can be formed by concatenating these stems and ending suffixes. For example, the verb-stem "  " (Latin-transliteration: "yabv") can be followed by more than 40 ending suffixes to generate the corresponding new words. Take noun-stem "" (Latin- transliteration: "sandali") as another example. This stem can be followed by 16 different ending suffixes and generate 16 new words. Table 1 partially lists the possible generated words for the two stems.

This work is supported by the Natural Science Foundation of China (NSFC) (NO. 61263037, NO. 71163029) and the Natural Science Foundation of Inner Mongolia of China (NO. 2011ZD11).

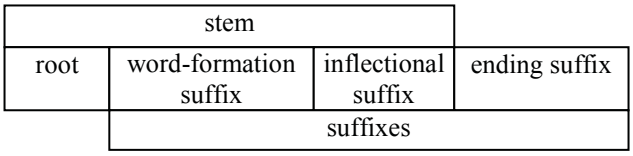

Fig. 1. The relationship of root, stem and suffixes of a word in Mongolian

3. THE SEGMENTATION-BASED MONGOLIAN LVCSR APPROACH

The basic idea of the proposed approach is as follows: we have noticed that the low coverage problem is caused by the combination of stem and suffixes and that the ending suffixes only affect the form but not the meaning. If the stems and ending suffixes can be recognized individually, only about 30000 stems and 400 suffixes are needed to be stored in the pronunciation dictionary. Therefore, stems and ending suffixes are adopted as the recognition unit for Mongolian.

According to this idea, the acoustic model and language model do not need to be changed much. However, some issues still need to be considered in the procedure for building them.

3.1 Building of the segmentation-based acoustic model

Before training the acoustic model, we first segmented the labeled text, which corresponds to the speech, to the stem and ending suffix form, and then converted them to the phoneme form for the training of the acoustic model. Each ending suffix has many pronunciations. In general, the first pronunciation in the pronunciation dictionary will be selected when converting the labeled text to its corresponding phoneme. This cannot guarantee the correctness of the selected pronunciation, thereby badly affects the precision of the trained acoustic model. In Mongolian, the pronunciation of an ending suffix is strongly related to the preceding stem under the constrains of Mongolian vowel harmony rules[4]. For example, the ending suffix "ᠭᠰᠠᠨ" (Latin-transliteration: "gsan") has six different pronunciations in total. But when the preceding stem is specified, its pronunciation will also be determined (see Table 2). Observing this fact, we proposed another technique which will select the most probable pronunciation rather than the first one by default.

Assume a stem S_1 is followed by an ending suffix S_2. Let $P(S_2)=\{p_1,p_2,...,p_m\}$ denote the possible pronunciations of S_2 and $C=\{c_1,c_2,...,c_n\}$ denotes the conditions of the Mongolian vowel harmony rules against which every Mongolian word needs to be tested. Then S_2's pronunciation PS_{S_2} can be determined by Formula (1) as

Table 1. The example of new words generated by Mongolian stem add ending suffix

Stem	Ending suffix	Generated words
ᠶᠠᠪᠤ (abv)	ᠬᠤ (hv)	ᠶᠠᠪᠤᠬᠤ (yabvhv)
	ᠵᠠᠢ (jai)	ᠶᠠᠪᠤᠵᠠᠢ (yabvjai)
	ᠨ᠋ᠠ (n_a)	ᠶᠠᠪᠤᠨ᠋ᠠ (yabvn_a)
	ᠶ᠋ᠠ (y_a)	ᠶᠠᠪᠤᠶ᠋ᠠ (yabvy_a)
	ᠯ᠋ᠠ (l_a)	ᠶᠠᠪᠤᠯ᠋ᠠ (yabvl_a)
	ᠮᠠᠭᠭᠠ (magqa)	ᠶᠠᠪᠤᠮᠠᠭᠭᠠ (yabvmagqa)
ᠰᠠᠨᠳᠠᠯᠢ (sandali)	ᠶᠢᠨ (-yin)	ᠰᠠᠨᠳᠠᠯᠢ ᠶᠢᠨ (sandali-yin)
	ᠳᠤ (-dv)	ᠰᠠᠨᠳᠠᠯᠢ ᠳᠤ (sandali-dv)
	ᠢ (-yi)	ᠰᠠᠨᠳᠠᠯᠢ ᠢ (sandali-yi)
	ᠠᠭᠠ (-aqa)	ᠰᠠᠨᠳᠠᠯᠢ ᠠᠭᠠ (sandali-aqa)
	ᠪᠠᠷ (-bar)	ᠰᠠᠨᠳᠠᠯᠢ ᠪᠠᠷ (sandali-bar)

Table 2. The different pronunciations of the ending suffix ᠭᠰᠠᠨ when following different stems

Ending suffix	Pronunciat-ions	Generated words	Pronunciation of words
ᠭᠰᠠᠨ (gsan)	sǎn	ᠶᠠᠪᠤᠭᠰᠠᠨ (yabvgsan)	jabsǎn
	Isǎn	ᠨᠠᠰᠤᠵᠢᠭᠰᠠᠨ (nasvjigsan)	nasdʒIgsǎn
	ǎsǎn	ᠵᠠᠯᠭᠠᠯᠳᠤᠭᠰᠠᠨ (jalgaldvgsan)	dʒalgǎldǎsǎn
	sǒn	ᠵᠤᠸᠢᠶᠠᠭᠰᠠᠨ (jwhiyagsan)	dʒœxœːsǒn
	Isǒn	ᠸᠠᠭᠢᠭᠰᠠᠨ (wqigsan)	ɔŋIsǒn
	ǒsǒn	ᠵᠤᠸᠭᠰᠸᠭᠰᠠᠨ (jwgswgsan)	dʒǒgsǒsǒn

follows:

$$PS_{S_2}=\arg\max_{p_j\in P(S_2)}\sum_{i=1}^{n}I(c_i,S_1,p_j) \quad (1)$$

where

$$I(c_i,S_1,p_j)=\begin{cases}1 & if\ S_1\ and\ p_j\ satisfies\ condition\ c_i\\ 0 & otherwise\end{cases} \quad (2)$$

In other words, knowing the stem S_1, we believe the true pronunciation of S_2 is the one which satisfies the most constraints of the Mongolian vowel harmony rules. Take the ending suffix "ᠭᠰᠠᠨ" (Latin-transliteration: "gsan") for example. When it's following the stem "ᠶᠠᠪᠤ", it's pronunciation will be "sǎn" ($\sum I$=2 is the maximum value among {2,1,1,1,0,0}), as shown in Table 3.

When training the acoustic model, we first built the context-dependent tri-phone model based on the

!START minu yarihv gejv baig_a ni tegun-u yarigsan enegu erge tegsi qinar-tv gadagadv guqun-v nulugelel-iyer vqaragvljv bugui jarim erge tegsi bvsv ujegdel bwlwn_a !END

The text before the segmentation

!START minu yari %hv ge %jv bai %g_a ni tegun %-u yari %gsan enegu erge tegsi qinar %-tv gadagadv guqun %-v nulugelel %-iyer vqaragvl %jv bugui jarim erge tegsi bvsv ujegdel bwl %wn_a !END

The text after the segmentation

Fig. 2. Comparison of the text before and after segmentation under the Latin transliteration perspective

Table 3. Choosing of the pronunciation of the ending suffix "ᠰᠠᠨ" when following the stem "ᠪᠠ" by adopting the Mongolian vowel harmony rules

Ending suffix	Pronunciat-ions	Conditions satisfied	$\sum I$
ᠰᠠᠨ gsan	sǎn	c_1 (the pronunciation of stem "ᠪᠠ" is of type "aevu") c_4 (the pronunciation of stem "ᠪᠠ" ends with consonant)	2
	Isǎn	c_1 (the pronunciation of stem "ᠪᠠ" is of type "aevu")	1
	ǎsǎn	c_1 (the pronunciation of stem "ᠪᠠ" is of type "aevu")	1
	sǒn	c_4 (the pronunciation of stem "ᠪᠠ" ends with consonant)	1
	Isǒn	None	0
	ǒsǒn	None	0

decision tree [2][5]. After that, we built a Continuous Hidden Markov Model (CHMM) [3][6] with Gaussian mixture distribution as the acoustic model.

3.2 Building of the segmentation-based language model

In Mongolian, a letter will have different visual forms in different words. Even in the same word, it still will be in different visual forms when placed at different locations (initial, middle and final). Moreover, the Mongolian letter for "ᠸ" (w) and "ᠸ" (v), "ᠣ" (o) and "ᠣ" (u) have the same visual form no matter where they are placed; and "ᠲ" (t) and "ᠳ" (d) have the same form only when placed at the initial or middle position; for "ᠧ" (h) and "ᠶ" (g) they will have the same form when used as feminine [4][7]. What's more, different pronunciation may be mapped to the same form of a letter. Considering this fact, we adopt the Latin-transliteration representation for both the labeled text and the text corpus, which will provide convenience for the following work.

Before training the language model, we first convert all the Mongolian words to the corresponding segmented forms, i.e., stem followed by ending suffix. Figure 2 gives an example of the comparison of the text before and after segmentation, where tokens begin with % denote ending suffixes. One constraint must be satisfied when segmenting the words is that the generated tokens must be legal stems or suffixes, otherwise the segmentation is not permitted.

Another issue must be taken into account is that some suffixes, when placed after the consonant stem, will have some additional vowels added before them. For example, when the stem "ᠠᠪ" (Latin-transliteration: "ab") is followed by the ending suffix "ᠮᠠᠭᠠ" (Latin-transliteration: "magqa"), there should be an additional vowel "ᠸ" (Latin-transliteration: "v") added and making the generated word to be "ᠠᠪᠸᠮᠠᠭᠠ" (Latin-transliteration: "abvmagqa"). To deal with these cases, we stored the augmented forms of these ending suffixes into the dictionary too. For example, for the ending suffix "ᠮᠠᠭᠠ", the other three forms, i.e. "ᠮᠠᠭᠠ" (Latin-transliteration: "magqa"), "ᠪᠮᠠᠭᠠ" (Latin-transliteration: "vmagqa") and "ᠸᠮᠠᠭᠠ" (Latin-transliteration: "wmagqa") will also be stored in the ending suffix dictionary.

The generated corpus are then used to train the N-gram language model [8]. The main advantage of using this training corpus is that it can not only enlarge the coverage of the words that can be recognized; but also, greatly alleviate the sample sparseness problem by increasing the number of occurrence and co-occurrence of the words.

Table 4. Datasets used in the Experiments

Datasets	# sentences	Length
Training	36000	about 50 hours
DIALOGUE	13000	about 13.5 hours
BOOK	5121	about 7.5 hours

4. EXPERIMENTS

We implement the LVCSR system based on the Hidden Markov Model Toolkit (HTK) [9] and use the SRI Language Modeling Toolkit (SRILM) [8] to train the language model. The training and testing sets are listed in Table 4, where the testing set is composed of two parts (DIALOGUE: recording of the Mongolian dialogues; BOOK: recording of Mongolian text books of junior school). For all of our experiments, we choose the context-dependent tri-phone Gaussian Mixture Model as the acoustic model. The mixture coefficient is set to 14.

To test the performance of the acoustic model built, we assume that all the words in the language model are uniformly distributed. Two groups of recognition experiments are performed: one of them has the ending

Table 5. Results achieved by the acoustic models with and without correction of Mongolian vowel harmony rules

	DIALOGUE		BOOK	
	WRR	PRR	WRR	PRR
NoLm-Seg	29.49%	74.11%	22.36%	67.99%
NoLm-Seg-Ad	**30.79%**	**75.83%**	**23.45%**	**69.24%**

Table 6. Performance of bigram and trigram Language Models for the segmented Mongolian words

	DIALOGUE(WRR)	BOOK(WRR)
2-gram-Seg	64.59%	55.40%
2-gram-Seg-Ad	65.03%	56.48%
3-gram-Seg	66.03%	58.72%
3-gram-Seg-Ad	**66.20%**	**59.74%**

suffix be corrected by the proposed correction method before training the acoustic model and the other does not. We used the Hvite in the HTK to perform decoding and adopted the widely used Word Correct Recognition Rate (WRR) and Phone Correct Recognition Rate (PRR) as our evaluation metrics. Table 5 lists the experimental results, from which we can observe that there is significant improvement if the pronunciation of the ending suffixes are corrected before the acoustic models are trained.

The performance of bigram and trigram language models for the segmented Mongolian words are also tested in our experiments. The results are listed in Table 6, where 2-gram-Seg denotes that the bigram language model is adopted. Similarly, 3-gram-Seg denotes that the trigram language model is adopted. The runs with "-Ad" mean that their acoustic models are trained by the text corrected by Mongolian vowel harmony rules. From Table 6, we can directly observed that the acoustic models with the training set corrected can greatly outperform that without correction and that the higher order models (3-gram) are better than the lower ones (2-gram).

We have also compared the performance of the segmentation-based LVCSR system (3-gram-Seg-Ad) with that of the whole word based (3-gram-Word-SR). The pronunciation dictionary for 3-gram-Word-SR is constructed from the Mongolian Orthography Dictionary [10] which contains 33918 words. For 3-gram-Seg-Ad, the pronunciation dictionary contains 35141 stems and 414 ending suffixes. We use the training and testing set listed in Table 4 once again to perform this test. According to our statistics, the pronunciation dictionary of 3-gram-Word-SR covers only 44.06% of the DIALOGUE words and 59.79% of the BOOK words. But for that of 3-gram-Seg-Ad, its coverage reaches nearly 100%. Figure 3 illustrates the experimental results, from which we can see that the WRR for 3-gram-Seg-Ad is much higher than that of 3-gram-Word-SR. More specifically, 35.6% improvement is achieved by the 3-gram-Seg-Ad system on the DIALOGUE testing set and 21.61% on the BOOK one.

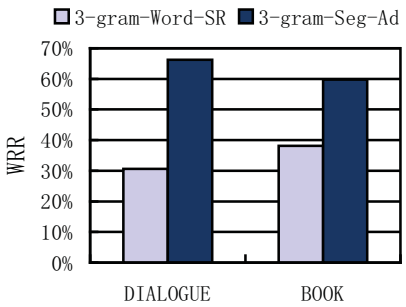

Fig. 3. Comparison between the WRR of the 3-gram-Word-SR and 3-gram-Seg-Ad

5. CONCLUSION

Based on the characteristic of the Mongolian word formation, we proposed a segmentation-based Mongolian LVCSR approach and depicted the details for building the acoustic model and the language model. There are three main contributions of our approach: first, the problem of low coverage of the pronunciation dictionary in the traditional Mongolian LVCSR system is nicely solved; secondly, the sample sparseness problem is greatly alleviated by increasing the number of occurrence of words; finally, it provides a new scheme to solve the similar problems in other agglutinative languages.

6. REFERENCES

[1] Guanglai Gao, Biligetu, Nabuqing, Shuwu Zhang, "A Mongolian Speech Recognition System Based on HMM", *International Conference on Intelligent Computing 2006(ICIC2006)*, Kunming, China, Aug. 2006, pp. 667-676.

[2] Qilao Hasi, Guanglai Gao, "Researching of Speech Recognition Oriented Mongolian Acoustic Model", *Chinese Conference on Pattern Recognition 2008 (CCPR 2008)*, Beijing, China, Dec. 2008, pp. 406-411.

[3] Feilong Bao, Guanglai Gao, "Improving of Acoustic Model for the Mongolian Speech Recognition System", in *Proc. Chinese Conference on Pattern Recognition 2009 (CCPR2009)*, Nanjing, China, Nov. 2009, pp. 616-620.

[4] Qingge'ertai, *Mongolian Syntax*, Inner Mongolia people publishing house , Hohhot, pp.77-133, 1991.

[5] Reichl W and Chou W. "Robust Decision Tree State Tying for Continuous Speech Recognition". *IEEE Trans Speech and Audio Processing*, vol. 8(5): 555-566, 2000.

[6] Jiqing Han, et al. *Speech Signal Processing*, Tsinghua University Publishing House, Beijing, pp. 200-218, 2004.

[7] Quejingzhabu, *Mongolian Codes*, Inner Mongolia University Publishing House, Hohhot, pp. 4-8, 2000.

[8] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", *International Conference Spoken Language Processing 2002,* Denver, Colorado, 2002. Vol. 2, pp. 901–904.

[9] Young S, et al. *The HTK book (Revised for HTK version 3.4.1)*, Cambridge University, 2009.

[10] Temusurvn and Otegen, *Mongolian Orthography Dictionary,* Inner Mongolia People Publishing House, Hohhot, 1999.